Prototype de Reconnaissance de Caractères Arabes Manuscrits à Base de Sous-Réseaux Neuronaux

L. Souici *
T. Sari *
Z. Zemirli **
M. Sellami *

Résumé

Dans ce papier, nous présentons un système connexionniste omni-scripteur et hors-ligne pour la reconnaissance de caractères arabes manuscrits. La reconnaissance de l'écriture arabe, pose, en plus des problèmes communs aux autres langues (disposition spatiale du texte, taille du vocabulaire et contraintes morphologiques dues aux variations propres au scripteur), d'autres problèmes spécifiques tels que la diversité des formes possibles d'une lettre d'après sa position dans un mot : [لبق قلب لقب قبل], laressemblance entre certains caractères qui ne diffèrent que par l'existence et l'emplacement d'une partie secondaire par rapport à la partie primaire [ب ت ث].

La méthode retenue au départ [SOU 96] adopte une approche intermédiaire entre les approches classiques basées sur l'extraction de paramètres et les méthodes neuronales pures, et ce, en utilisant un réseau neuronal (Perceptron Multicouche) alimenté par des caractéristiques (métriques, statiques et structurelles) extraites à partir des images des caractères scannérisés. Le prototype conçu a été développé en exploitant l'environnement NeuroShell, Neuro- Windows (Ward Systems Group, Inc) .

Pour améliorer les résultats obtenus par la première approche qui consistait à reconnaître les caractères indépendamment de leur forme et leur position dans un mot ([SEL 97], [SOU 97]), nous avons opté pour une nouvelle architecture basée sur un découpage en quatre sous-réseaux, dédiés chacun à l'apprentissage d'une forme pour l'ensemble des caractères (isolé, au début, au milieu, et à la fin d'une composante connexe). Ainsi, lors de l'identification, un algorithme de détection de ligature (attache ou lien) a été mis en oeuvre dans le but d'acheminer le caractère vers le sous réseau approprié.

A l'issue des étapes d'apprentissage et de test, et, dans le but d'en accroître les performances, un post-traitement est effectué afin de renforcer les résultats fournis par le système connexionniste couplé à un système à base de règles pour la détection de chaînes inconsistantes. L'analyse morphologique des chaînes (consistantes) obtenues permet le passage à la reconnaissance de texte, l'aspect sémantique n'étant pas pris en compte pour le moment.

L'objectif à moyen terme serait d'envisager d'élargir l'utilisation de ce prototype à des applications qui faciliteraient l'automatisation du traitement d'informations manuscrites (registres d'état civil, actes notariés) ainsi qu'à des applications à caractère pédagogique.

Mots-Clés : *Reconnaissance de caractères arabes manuscrits, Réseaux neuronaux, Apprentissage, Détection de ligatures, Analyse morpho-lexicale.*

1. INTRODUCTION

La lecture optique des textes et la reconnaissance d'écriture manuscrite ont été des domaines de recherche actifs surtout durant ces dernières années ([LOR 92], [MOR 92]). Si pour des textes imprimés ou dactylographiés, les principales difficultés ont été surmontées, la situation est complètement différente en ce qui concerne la reconnaissance des textes manuscrits. Les problèmes rencontrés sont liés essentiellement à la disposition spatiale du texte, le nombre de scripteurs, la taille du vocabulaire ainsi que les contraintes dues aux variations et déformations propres au scripteur, relatives aux conditions d'écriture ou au contexte lexical.

Les réseaux neuronaux, grâce à leur non-linéarité et à leur grande capacité d'apprentissage, se sont montrés très compétitifs dans le domaine de la reconnaissance de caractères ([JAI 96], [HAR 93], [YAN 94]).

La recherche dans le domaine de la reconnaissance des caractères arabes n'a débuté que tardivement ([ABU 94], [ALB 95], [ALM 87], [ALT 95], [ALY 92], [AMI 96], [AUD 93]). Ceci étant dû, entre autres, aux particularités spécifiques à l'écriture arabe.

En utilisant l'approche connexionniste, nous nous sommes intéressés à la reconnaissance de texte arabe manuscrit, le système projeté se présente selon la Figure 1. Dans ce papier, nous détaillerons les points suivants :

- Description du prototype de reconnaissance de caractères avec les modules qui le constituent et des exemples de résultats obtenus après son utilisation..
- Description du post-traitement effectué pour la reconnaissance de mots grâce à un système à base de règles (détection de chaînes inconsistantes) et à un analyseur morphologique.

2. PROTOTYPE DE RECONNAISSANCE DE CARACTERE

Le prototype proposé présente l'architecture typique d'un système de reconnaissance de caractères composé des modules d'acquisition, de segmentation, d'extraction de caractéristiques, d'apprentissage et de test.

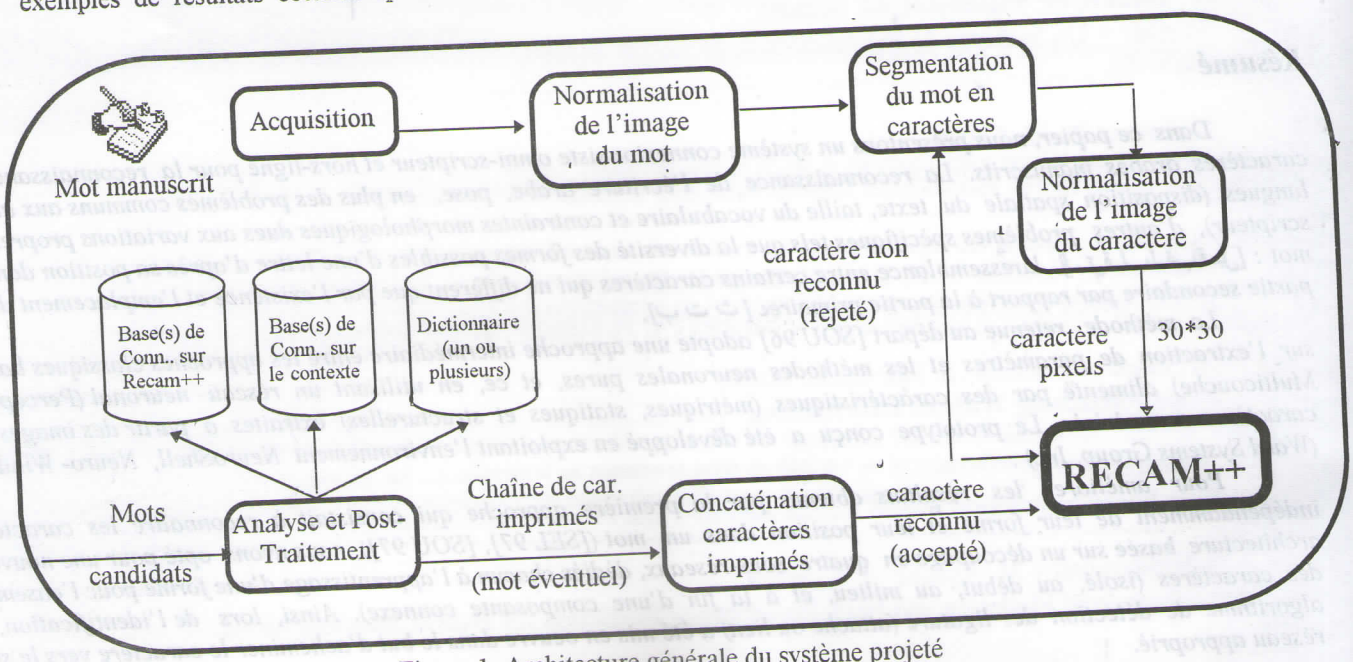

Figure 1: Architecture générale du système projeté

2.1. Acquisition

Un scanner de résolution 600 dpi, noir et blanc avec 16 niveaux de gris permet de transformer les échantillons de caractères manuscrits (utilisés pour entraîner et tester le prototype) en images ayant le format BMP (opération de digitalisation ou de binarisation). Des échantillons de mots séparés par des blancs ont été recueillis auprès de plusieurs scripteurs sur des documents précasés.

2.2. Segmentation et prétraitement

Les images des caractères ont été normalisées à une taille de 30?30 pixels. La phase de segmentation peut être éliminée car la présence préalable de cases sur les imprimés utilisés pour l'acquisition, oblige l'utilisateur (discipliné) à segmenter son texte en l'écrivant.

2.3. Extraction de Caratéristiques

Les paramètres que nous avons retenus sont ([BUR 92], [SEL 97],[SOU 97]) :

- **Caractéristiques Métriques**: représentent les profils normalisés des caractères. Nous avons 60 paramètres, 30 pour le profil droit et 30 pour le profil gauche.
- **Caractéristiques Statiques**: ces paramètres caractérisent la distribution des pixels à l'intérieur du cadre divisé en régions de 6 manières différentes (6 subdivisions). On obtient un total de 19 régions (19 caractéristiques) :

$$Rj = Nj / N.$$

Nj: nombre de pixels noirs à l'intérieur de la région j.
N : nombre total de pixels dans le caractère.

- **Caractéristiques Structurelles:** elles permettent d'avoir plus d'informations sur le type de caractère considéré. On distingue deux types de projections: 30 pour les projections horizontales et 30 pour les projections verticales.

L'ensemble de ces 139 caractéristiques (60+19+60) vont représenter les entrées du réseau neuronal.

2.4. Apprentissage Neuronal

Le noyau du système est un Perceptron Multi-Couches ayant une seule couche d'entrée (caractéristiques extraites), puis décomposé en quatre sous-réseaux constitué chacun d'une couche cachée et d'une couche de sortie (Figure 3). Pour chaque sous réseau, la couche cachée comporte 4 groupes de neurones (slabs) dont le nombre est choisi avec des heuristiques puis affiné selon les résultats obtenus :

1er groupe : 17 neurones
2ème groupe : 17 neurones
3ème groupe : 13 neurones.
4ème groupe : 13 neurones

La couche de sortie désigne l'ensemble de caractères à reconnaître, elle comporte 4 groupres de neurones :

1er groupe	: 16 neurones	isolés
2ème groupe	: 16 neurones	finaux
3ème groupe	: 10 neurones	milieu
4ème groupe	: 10 neurones	début

Le premier sous réseau SR1 est utilisé pour la reconnaissace .des caractères arabes manuscrits isolés, le 2ème sous réseau SR2 pour les caractères finaux, les caractères qui s'écrivent au milieu ou au début d'un mot sont pris en charge respectivement par les sous réseaux 3 et 4 (SR3 et SR4). Soit l'exemple suivant :

MOT	SR 1 Isolé	SR 2 Fin	SR 3 Milieu	SR 4 Début
علوم	م	و	لـ	ع
قناع	ع	ا	ن	ق

L'apprentissage des sous-réseaux s'est effectué séparément pour chaque sous-ensemble représentatif d'une classe de caractères. Le plus important dans cette phase est de savoir quand il faùt arrêter l'entraînement tout en s'assurant que le sous réseau est capable de généraliser en évitant les problèmes causés par l'apprentissage (surapprentissage, oubli catastrophique) [FRE 92], [JAI 94], [JOD 94], [RUM 86], [SIM 90].

C'est donc à partir de cette étape que le système doit ajuster ses paramètres afin de donner une réponse **lors de la phase de test.**

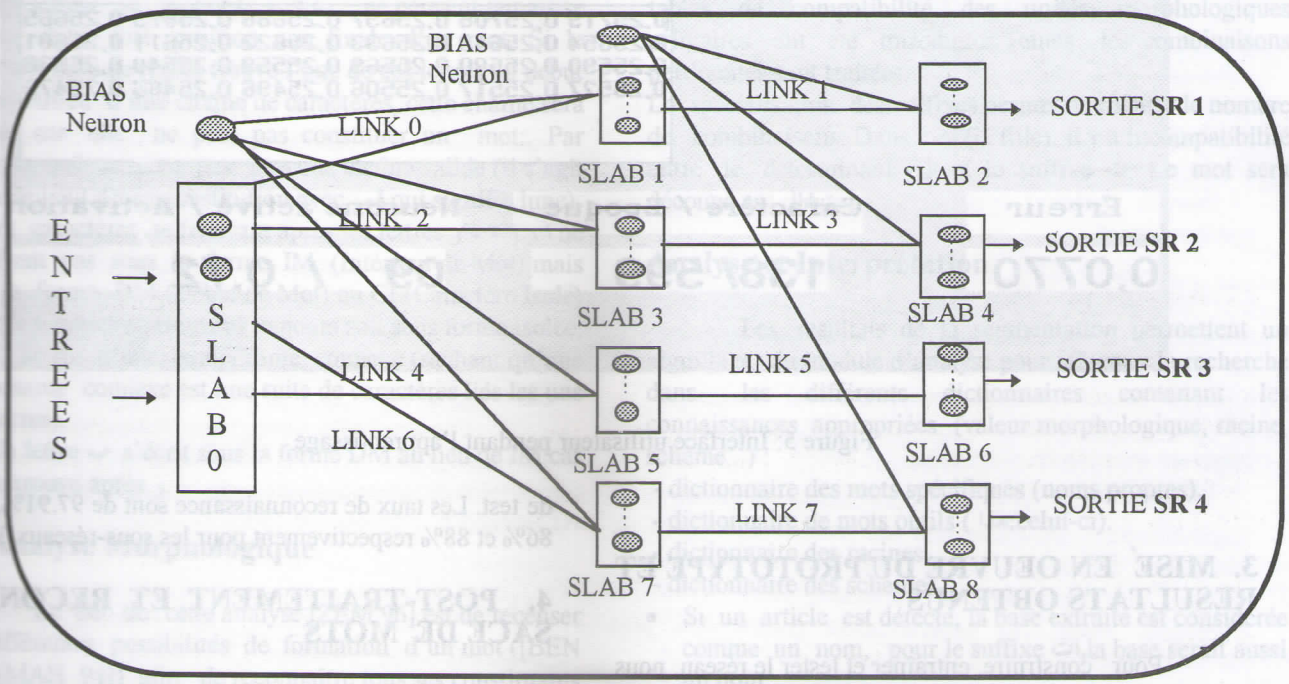

Figure 3: Architecture générale du réseau

2.5. Test du Réseau

Le test permet d'évaluer la capacité de généralisation du réseau sur des exemples non appris (caractères écrits par d'autres scripteurs). On procède à l'extraction des caractéristiques les concernant, puis on lance le test. Cette phase ne peut avoir lieu qu'après le chargement d'un réseau déjà entraîné sur les caractères écrits séparément sous leurs différentes formes.

La notion de classification intervient à ce stade: en plus de l'extraction des caractéristiques, la classe à laquelle appartient le caractère doit être identifiée mais le problème qui se pose est de déterminer vers quel sous réseau seront propagées les entrées?

C'est justement, la détection de ligatures qui résoud ce problème en identifiant le sous réseau approprié et permettre ainsi la propagation.

Pour cela , nous avons utilisé une procédure inspirée d'une méthode de détection de contours [MOR 92]. [NAM 90].

Exemple illustratif

1 2 3 4

Figure 4 : Détection de ligatures pour la répartition.

1-Aucun lien (caractère isolé).
2-Présence d'un lien droit, absence de lien gauche (caractère en fin de mot).
3-Absence de lien droit, présence de lien gauche (caractère au début de mot).
4-Présence d'un lien droit et d'un lien gauche (caractère au milieu de mot).

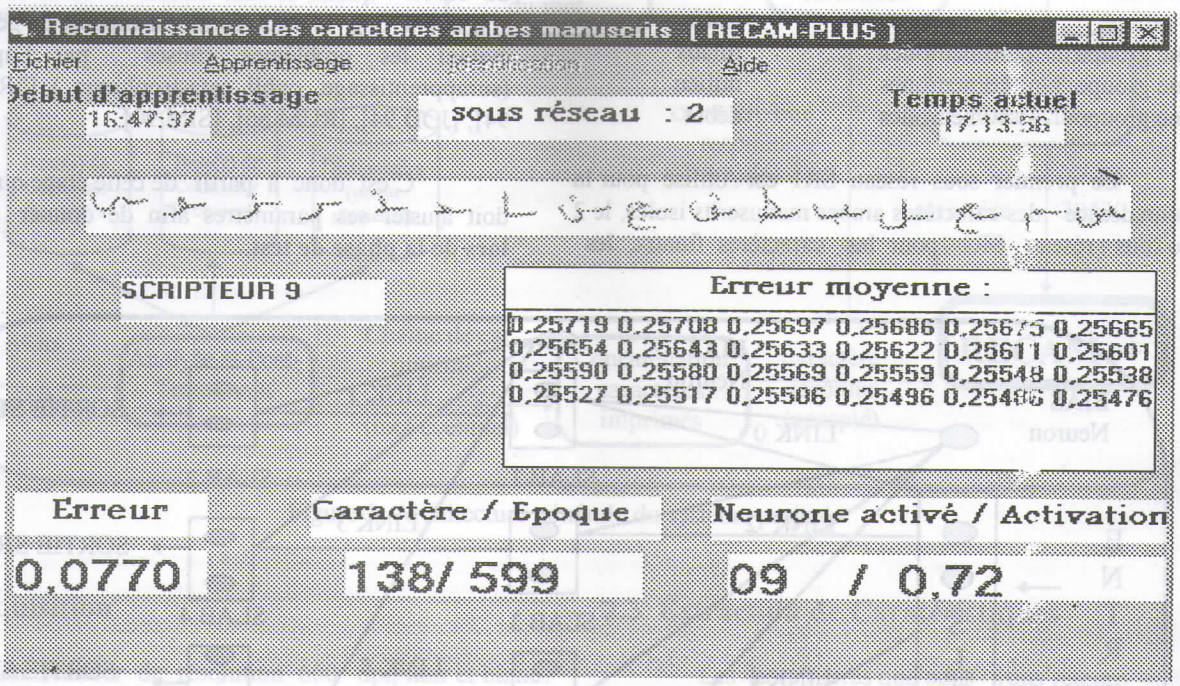

Figure 5: Interface utilisateur pendant l'apprentissage

3. MISE EN OEUVRE DU PROTOTYPE ET RESULTATS OBTENUS

Pour construire, entraîner et tester le réseau, nous avons exploité NeuroWindows (de Ward Systems Group, Inc [NEU 95]) qui est une bibliothèque dynamique de fonctions (Dynamic Link Library: DLL) conçue pour être utilisée avec les langages de programmation Visual Basic (ou Access Basic) de Microsoft.

Pour la mise en oeuvre du système, nous avons développé un ensemble de procédures concernant l'extraction de caractéristiques, la répartition vers le sous-réseau approprié ainsi que les processus d'apprentissage et de test. Les taux de reconnaissance sont de 97.91%, 97.5%, 86% et 88% respectivement pour les sous-réseaux 1, 2, 3 et

4. POST-TRAITEMENT ET RECONNAISSACE DE MOTS

Après avoir testé la capacité de reconnaissance du réseau, des ensembles de mots écrits par des scripteurs (ayant participé ou non aux étapes précédentes) sont présentés au prototype pour effectuer la reconnaissance des caractères qui les constituent. Chaque chaîne de caractères imprimés délivrée en résultat représente une entité lexicale qu'il faudra analyser et traiter morphologiquement afin d'en déterminer les différents composants et de vérifier leur validité ainsi que leur compatibilité.

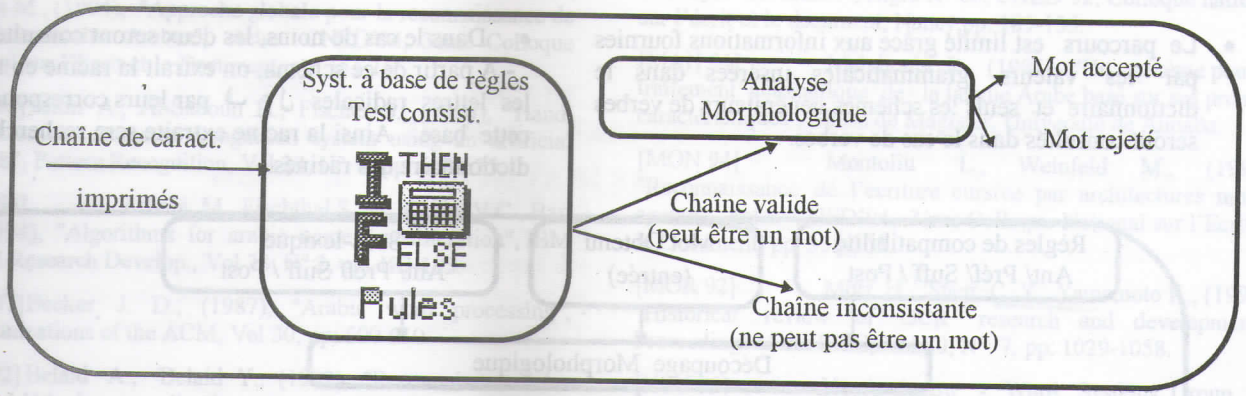

Figure 6: Processus de reconnaissance de mots

4.1. Détection de Chaînes Inconsistantes

Un système à base de règles vérifie si la combinaison de caractères obtenus est acceptable ou inconsistante ([ARB 94]) pour décider s'il s'agit ou non d'un mot valide en appliquant un ensemble de règles relatives à la position du caractère dans le mot (CI: caractère isolé, DM: début de mot, IM: intérieur de mot, FM: fin de mot), ainsi qu'à des restrictions sur l'utilisation de certaines formes ou de certaines combinaisons de caractères :

- Les lettres ه ة ـة ى ي ـى ـي ne peuvent jamais se trouver dans une position non finale d'un mot. Si le système à base de règles trouve l'une de ces lettres au début ou au milieu d'une chaîne de caractères, cette chaîne sera rejetée car elle ne peut pas constituer un mot:. Par exemple قمر ne peut pas être une chaîne valide (il s'agit probablement d'une substitution avec قمر qui signifie lune)
- Les caractères se trouvant après les lettres و ذ د ز ر ne s'écrivent pas sous la forme IM (Intérieur de Mot) mais sous la forme DM (Début de Mot) ou CI (Caractère Isolé) car ces 6 lettres se trouvent toujours soit sous forme isolée, soit à la fin d'une composante connexe (sachant qu'une composante connexe est une suite de caractères liés les uns aux autres).

رب: la lettre ب s'écrit sous la forme DM au lieu de IM car elle se trouve après ر.

4.2.Analyse Morphologique

Le but de cette analyse [ZEM 96] est de recenser les différentes possibilités de formation d'un mot ([BEN 85], [MAH 94]) afin de reconnaître tous ses constituants ([SAR189], [SAR289], [SOU 89]) pour les traiter convenablement et de prévoir une bonne organisation du lexique qui ne doit contenir que les formes de base des mots ainsi que la manière d'accéder à ce lexique. L'analyse d'un mot passe par les étapes suivantes (voir Figure 7) :

- Découpage morphologique d'un mot en cinq éventuelles parties (de droite à gauche) :

[Postf.]+[Suf.]+[Radical nom. ou verb.]+[Préf.]+[Antéf.]

avec vérification de l'appartenance des affixes (postfixes, suffixes, préfixes et antéfixes) au lexique et de leur compatibilité entre eux..

- Analyse selon le découpage choisi pour l'accès aux informations lexicales.
- Interprétation des résultats obtenus.

Segmentation

Des règles de décomposition représentées par des tables de compatibilité des unités morphologiques primaires ont été introduites.Toutes les combinaisons pertinentes sont traitées.

La spécialisation des affixes permet de réduire le nombre de combinaisons: Dans البنت(la fille), il y a incompatibilité entre le déterminant ال et le suffixe ت. Le mot sera découpé en: بنت+ال.

Analyse et Interprétation

Les résultats de la segmentation permettent un aiguillage du module d'analyse pour effectuer la recherche dans les différents dictionnaires contenant les connaissances appropriées (valeur morphologique, racine, schème...) :

- dictionnaire des mots spécifiques (noms propres).
- dictionnaire de mots outils (هذا:celui-ci).
- dictionnaire des racines.
- dictionnaire des schèmes.

- Si un article est détecté, la base extraite est considérée comme un nom, pour le suffixe ات,la base serait aussi un nom.
- Si le mot n'est pas affixé, la recherche est lancée dans le dictionnaire des mots spéciaux (et des mots outils), dans le cas contraire, on teste la compatibilité entre les mots spéciaux (outils) et l'affixe détecté.

Avant de trouver la racine d'une base, il faut déterminer son schème, en consultant le dictionnaire des schèmes classés par longueur.

- Le parcours est limité grâce aux informations fournies par les valeurs grammaticales insérées dans le dictionnaire et seuls les schèmes générateurs de verbes seront consultés dans le cas de verbes.
- Dans le cas de noms, les deux seront consultés.

- A partir de ce schèma, on extrait la racine en substituant les lettres radicales ف ع ل par leurs correspondants dans cette base. Ainsi la racine extraite sera recherchée dans le dictionnaire des racines.

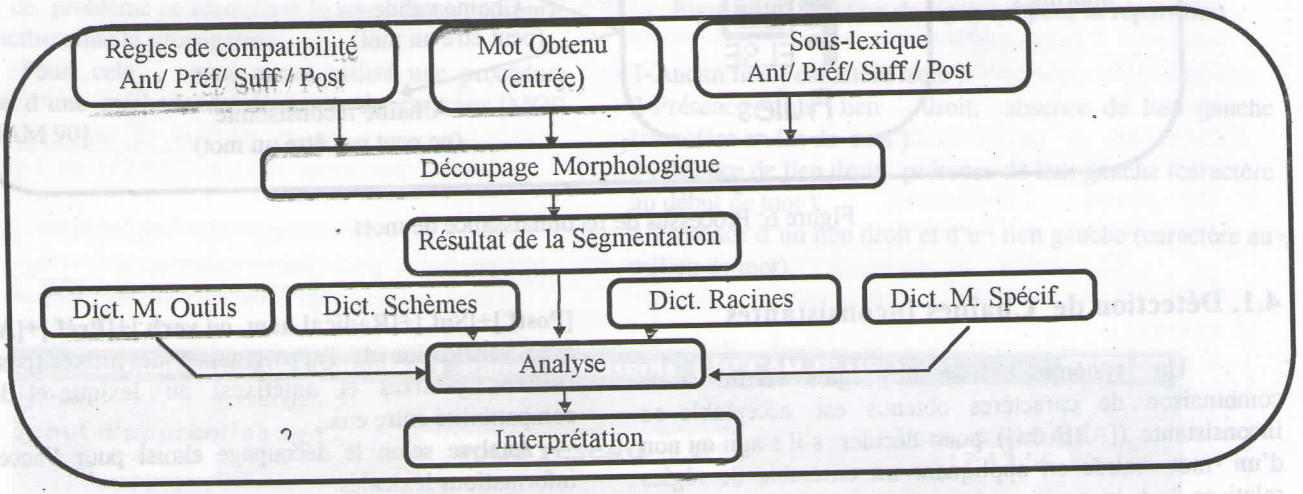

Figure 7: architecture générale de l'analyseur morphologique

5.CONCLUSION ET PERSPECTIVES

Le choix de la nouvelle architecture basée sur les sous réseaux a pour avantage de:

- Faciliter l'élargissement de la base d'apprentissage en nombre de scripteurs et de caractères à reconnaître en permettar t l'apprentissage des sous-réseaux séparément.
- Minimiser les risques d'ambigüité inter-classes.
- Pouvoir exploiter le parallélisme en cas de présence de système réparti.

L'application dans son ensemble est en cours d'évaluation, les tests effectués nous ont fait prendre conscience de la complexité des modèles connexionnistes, de par la multitude des objets qu'ils manipulent et de leurs effets sur les performances (temps de calcul, occupation mémoire et fiabilité). A moyen terme, nous prévoyons plusieurs possibilités d'évolution:

- Elargir encore la base d'apprentissage du prototype de reconnaissance de caractères en introduisant plus de caractères et un plus grand nombre de scripteurs pour inclure un maximum de variations dans les styles d'écriture, ainsi, la généralisation pourra être plus performante.
- Prendre en considération les différentes combinaisons spéciales de deux lettres liées (لا) et les faire apprendre au réseau comme s'il s'agissait d'une seule lettre.
- Utiliser des paramètres supplémentaires (extraction de nouvelles caractéristiques) pour prendre des décisions finales concernant les caractères ambigus ainsi que dans le cas de chaînes inconsistantes rejetées par le système à base de règles.
- Utiliser encore d'autres architectures, règles d'apprentissage (non supervisées par exp) puis appliquer le principe de "*vote majoritaire*" qui permet de prendre une décision finale en se basant sur les réponses individuelles de chaque réseau: le système les collecte toutes et, dans le cas de différences choisit la plus fréquente.
- Introduire des connaissances syntaxiques et pragmatiques pour renforcer la reconnaissance de mots et passer au stade de reconnaissance de phrases. Remarquons que l'une des difficultés principales de la langue arabe est qu'il est possible qu'une chaîne de caractères ne contenant pas de délimiteur ne soit pas un mot mais toute une phrase, comme par exemple: سأفهمكه (signifiant: "Je vais te l'expliquer") et سألتمونيه(signifiant: "vous me l'avez demandé").

Ces exemples permettent d'illustrer les problèmes morphologiques pouvant être rencontrés par un système complet de reconnaissance de phrases écrites en langue Arabe.

REFERENCES

[ABU 94] Abuhaiba I.S.I., Mahmoud S.A., Green R.J, (1994) "Recognition of handwritten cursive Arabic characters", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 16, pp: 664-672.

[ALB 95]Al Badr B., Mahmoud S. A., (1995), "Survey and bibliography of arabic optical text recognition", Signal processing, Vol 41, pp: 49-77.

[ALY 92] Al-Yousefi H., Udpa S.S., (1992), "Recognition of Arabic characters", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 14, pp: 853-857.

[AME 94] Ameur A., Romeo-Pakker K., Miled H., Cheriet M., (1994), "Approche globale pour la reconnaissance de mots manuscrits Arabes", Actes CNED'94, 3ème Colloque National sur l'Ecrit et le Document, pp: 151-156.

[AMI 96] Amin A., Al-Sadoun H., Fischer S., (1996), "Hand-printed arabic character recognition system using an artificial network", Pattern Recognition, Vol 29, N°: 4, pp: 663-675.

[ARB 94] Arbabi M., Fischthal S. M., Cheng V.C., Bart E., (1994), "Algorithms for arabic name transliteration", IBM Journal Research Develop., Vol 38, N°:2, pp: 183-193.

[BEC 87] Becker J. D., (1987), "Arabic word processing", Communications of the ACM, Vol 30, pp: 600-610.

[BEL 92] Belaid A., Belaid Y., (1992), "Reconnaissance des formes: Méthodes et applications", InterEditions.

[BEN 85] Benhamouda B., (1985), "Morphologie et syntaxe de la langue arabe", Editions SNED, Alger.

[BUR 92] Burel G., Pottier I., Catros J.Y., (1992), "Recognition of handwritten digits by image processing and neural network", IEEE, IJCNN International Joint Conference on Neural Networks, Vol 3, pp: 666-671.

[FRE 92] Freeman J. A., Skapura D. M. , (1992), "Neural networks: Algorithms, applications and programming techniques", Addison Wesley Publishing Company.

[FUC 93] Fuchs C., (1993), "Linguistique et traitement automatique des langues", Chap 3, 4, 5, 8 (Morphologie, Syntaxe, Sémantique, Compréhension automatique des textes), HU Linguistique Hachette Supérieur.

[HAR 90] Harmalkar S., Sinha R. M. K., (1990), "Integrating word level knowledge in text recognition",Proceedings of ICPR'90, 10 th International Conference on Pattern Recognition, Vol 1, pp: 758-760.

[HAR 93] Harrienhausen-Muhlbauer B., Koop A., (1993), "SCRIPT: A prototype for the recognition of continuous, cursive, handwritten input by means of a neural network simulator", IEEE Intern. Conference on Neural Networks, Vol 3, pp: 1672-1677.

[HAR 95] Harrienhausen-Muhlbauer B., (1995), "SCRIPT-a system for the recognition of handwritten input using linguistic and statistical filter mechanism as well as crossword lexicon", 3èmes Journées internationales d'Analyse statistique des Données Textuelles, JADT'95, Italie, pp: 173-180.

[HOC 93] Hoch R., Kieninger T., (1993), "On virtual partitioning of large dictionaries for contextual post-processing to improve character recognition", IEEE, pp : 226-231.

[JAI 96] Jain A. I., Jianchiang Mao, Mohiuddin K.M., (1996), "Artificial neural netwoks: A tutorial ", Computer, Vol 29, N°: 3, pp: 31-44.

[JOD 94] Jodouin J. F. , (1994), "Les réseaux neuromimétiques: Modèles et applications", Hermès.

[LAL 95] Lallich-Boidin G., Rouault J., "Coopération statistique-linguistique pour l'analyse textuelle", 3èmes Journées internationales d'Analyse statistique des Données Textuelles, JADT'95, Italie, pp: 45-54.

[LOR 92] Lorette G, Lecourtier Y. , (1992), "Reconnaissance et interprétation de textes manuscrits hors-ligne: un problème d'analyse de scène?", Bigre N° 80, CNED'92, Colloque national sur l'écrit et le document, Nancy,pp: 109-135.

[MAH 94] Mahdjoubi R., (1994), "Un système pour le traitement automatique de la langue Arabe basé sur ses propres caractéristiques", Thèse de Magister, Université de Annaba.

[MON 94] Montoliu L., Weinfeld M., (1994), "Reconnaissance de l'écriture cursive par architectures multi-agents", Actes CNED'94, 3ème Colloque National sur l'Ecrit et le Document, pp: 355-363.

[MOR 92] Mori H., Suen C. Y., Yamamoto K., (1992), "Historical review of OCR research and development", Proccedings of the IEEE, Vol 80, N°: 7, pp: 1029-1058.

[NEU 95] Neurowindow - Ward Systems Group Inc (1995) "Neural Network Products for the next century". Executive Park West - Frederick, MD 21702 - USA

[RUM 86] Rumelhart D, Hinton G., Williams R. J. , (1986), "Learning internal representations by error propagations", In Parallel Distributed Processing, Explorations in the microstructure of cognition, Vol 1: Foundations, pp: 318-362, MIT Press.

[SAR189] Saroh A., Brusset J., (1989), "Un modèle de générateur morphologique de l'arabe", Actes RFIA'89, Congrès Reconnaissance des Formes et Intelligence Artificielle, Tome 1, pp: 263-270.

[SAR289] Saroh A., (1989), "Base de données lexicales dans un système d'analyse morpho-syntaxique de l'arabe SYAMSA", Thèse de Doctorat, UPS Toulouse.

[SEL 97] Sellami M., Souici L., Zemirli Z., (1997), "Système hybride pour la reconnaissance de l'arabe manuscrit", Colloque: le monde arabe et la société de l'information, Tunis, Mai 1997.

[SIM 90] Simpson P.K. , (1990), "Artificial neural systems : Foundations, paradigms, applications and implementations", Pergamon Press.[SOU 89] Souilem D., (1989), "Un système assisté par ordinateur pour la grammaire arabe SEAGA", Thèse de Doctorat, UPS Toulouse.

[SOU 96] Souici L., Farah N., Sellami M., (1996), "Contribution to the recognition of hand-written arabic text by a neural network in cooperation with an expert system ", Current trends in computer science and information systems, Philadelphia University, Jordan, pp: 13-23, July 1996.

[SOU 97] Souici L., Zemirli Z., Sellami M. , (1997), "Système connexionniste pour la reconnaissance de l'arabe manuscrit", JST'97-Francil, Avignon, France, pp: 383-388, Avril 1997.

[WEL 90] Wells C. J., Evett L. J., Whitby P. E., Whitrow R. J., (1990), "Fast dictionary look-up for contextual word recognition", Pattern Recognition, Vol 23, N°: 5, pp: 501-508.

[YAN 94] Yanikoglu B. A., Sandon P. A., (1994), "Recognizing off-line cursive handwriting", Proceedings IEEE Computer Society Conference on Computer Vision and Pat. Rec. , pp: 397-403.

[ZEM 96] Zemirli Z., N.Vigouroux., (1996), "Un analyseur morphologique destiné à l'aide à la constructiond'une base de données lexicales de la langue arabe", IERA'96, Rabat.

* **Institut d'Informatique - Université Annaba**

** **I.N.I. - M Oued-smar 16270 Alger Algérie**